International School of Economics at TSU

Econometrics II

Lasha Chochua

Problem Set 4

Instructions: You are encouraged to solve the problems before the recitation. Additionally, you are encouraged to work in groups. It is **not mandatory** to submit solutions unless stated otherwise. However, if you would like to share your solution, I would be happy to review it.

Problem 1: Consider the binary choice latent variable model

$$Y_i^* = X_i'\beta + e_i, \qquad Y_i = \mathbb{1}\{Y_i^* > 0\}, \qquad e_i \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

Suppose $\sigma^2 \neq 1$, but you estimate the standard probit specification that imposes unit error variance.

a. Show that the probit response probability under (1) is $\Phi(X_i'\beta/\sigma)$, and conclude that what the probit MLE actually identifies is $\beta^* = \beta/\sigma$, not β itself.

Solution

Compute directly:

$$\mathbb{P}(Y_i = 1 \mid X_i) = \mathbb{P}(X_i'\beta + e_i > 0 \mid X_i) = \mathbb{P}(e_i > -X_i'\beta \mid X_i).$$

Since $e_i/\sigma \sim \mathcal{N}(0,1)$ and the standard normal is symmetric around zero,

$$\mathbb{P}(e_i > -X_i'\beta \mid X_i) = \mathbb{P}\left(\frac{e_i}{\sigma} > -\frac{X_i'\beta}{\sigma}\right) = 1 - \Phi\left(-\frac{X_i'\beta}{\sigma}\right) = \Phi\left(\frac{X_i'\beta}{\sigma}\right).$$

The probit log-likelihood is identical for any pair (β, σ) that gives the same ratio β/σ. Therefore the data only pin down $\beta^* = \beta/\sigma$ – the original (β, σ) pair is not separately identified. The standard normalization $\sigma = 1$ is a labelling choice, not an empirical assumption.

b. Show that the marginal effect of X_{ij} on $\mathbb{P}(Y_i = 1 \mid X_i)$ is invariant to the normalization in the sense that the *true* marginal effect equals the marginal effect computed using the rescaled coefficient β^*.

Solution

The true marginal effect of X_{ij} on the response probability is, by the chain rule applied to $\Phi(X_i'\beta/\sigma)$,

$$\frac{\partial}{\partial X_{ij}}\Phi\left(\frac{X_i'\beta}{\sigma}\right) = \phi\left(\frac{X_i'\beta}{\sigma}\right)\cdot\frac{\beta_j}{\sigma} = \phi(X_i'\beta^*)\cdot\beta_j^*.$$

This is exactly the marginal effect formula that probit reports using the estimated $\hat{\beta}^*$. Although the unscaled β_j and σ are individually unidentified, their ratio $\beta_j^* = \beta_j/\sigma$ is identified, and the marginal effect depends only on this ratio. This is why **average marginal effects are reported** for substantive interpretation, while raw coefficient magnitudes have no intrinsic meaning.

Problem 2: Let $\hat{\beta}$ be the MLE for the logit model with i.i.d. data $\{(Y_i, X_i)\}_{i=1}^n$, $Y_i \in \{0,1\}$. Define the score and the negative Hessian as

$$S_n(\beta) = \sum_{i=1}^n X_i(Y_i - \Lambda(X_i'\beta)), \qquad \mathcal{H}_n(\beta) = \sum_{i=1}^n X_i X_i'\,\Lambda(X_i'\beta)(1-\Lambda(X_i'\beta)). \tag{2}$$

a. Derive the asymptotic distribution of $\hat{\beta}$ under correct specification, including an explicit derivation via a first-order Taylor expansion of the score around the pseudo-true value β_0.

Solution

The MLE solves the first-order condition $S_n(\hat{\beta}) = 0$. Taylor-expand $S_n(\hat{\beta})$ around the pseudo-true value β_0:

$$0 = S_n(\hat{\beta}) = S_n(\beta_0) + \frac{\partial S_n(\beta_0)}{\partial \beta'}(\hat{\beta}-\beta_0) + R_n,$$

where R_n is a remainder term that is negligible in large samples. The derivative of the score is

$$\frac{\partial S_n(\beta_0)}{\partial \beta'} = -\sum_{i=1}^n X_i X_i'\,\Lambda(X_i'\beta_0)(1-\Lambda(X_i'\beta_0)) = -\mathcal{H}_n(\beta_0).$$

Substitute and solve for $\sqrt{n}(\hat{\beta}-\beta_0)$:

$$\sqrt{n}(\hat{\beta}-\beta_0) = \left(\frac{1}{n}\mathcal{H}_n(\beta_0)\right)^{-1}\cdot\frac{1}{\sqrt{n}}S_n(\beta_0) + o_p(1).$$

By the law of large numbers,

$$\frac{1}{n}\mathcal{H}_n(\beta_0) \xrightarrow{p} Q \equiv \mathbb{E}[X_i X_i'\,\Lambda(X_i'\beta_0)(1-\Lambda(X_i'\beta_0))].$$

For the score, note that under correct specification $\mathbb{E}[Y_i \mid X_i] = \Lambda(X_i'\beta_0)$, so

$$\mathbb{E}[X_i(Y_i - \Lambda(X_i'\beta_0))] = 0.$$

The score is a sum of mean-zero i.i.d. terms with covariance

$$\Omega = \mathbb{E}[X_i X_i'(Y_i - \Lambda(X_i'\beta_0))^2] = \mathbb{E}[X_i X_i' \Lambda(X_i'\beta_0)(1 - \Lambda(X_i'\beta_0))] = Q,$$

where the second equality uses $\mathbb{E}[(Y_i - \Lambda_i)^2 \mid X_i] = \Lambda_i(1 - \Lambda_i)$ – the Bernoulli variance. So the **information matrix equality** $\Omega = Q$ holds. By the central limit theorem,

$$\frac{1}{\sqrt{n}} S_n(\beta_0) \xrightarrow{d} \mathcal{N}(0, Q).$$

By Slutsky's theorem,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, Q^{-1} Q\, Q^{-1}) = \mathcal{N}(0, Q^{-1}).$$

The sandwich variance $Q^{-1}\Omega Q^{-1}$ collapses to the simple inverse Hessian formula Q^{-1}.

b. Now suppose the model is misspecified: $\mathbb{P}(Y_i = 1 \mid X_i) = m(X_i)$ for some function m that does *not* equal $\Lambda(X_i'\beta)$ for any β. Show that the information matrix equality fails and that valid inference requires the full sandwich variance. Give an explicit expression for Ω under misspecification.

Solution

Under misspecification, the pseudo-true value β_0 is defined by the population first-order condition

$$\mathbb{E}[X_i(Y_i - \Lambda(X_i'\beta_0))] = 0,$$

not by $m(X_i) = \Lambda(X_i'\beta_0)$. Now compute the conditional variance of Y_i around its true mean:

$$\mathrm{Var}(Y_i \mid X_i) = m(X_i)(1 - m(X_i)) \neq \Lambda(X_i'\beta_0)(1 - \Lambda(X_i'\beta_0)).$$

The score variance becomes

$$\Omega = \mathbb{E}[X_i X_i'(Y_i - \Lambda(X_i'\beta_0))^2] = \mathbb{E}\left[X_i X_i'((m(X_i) - \Lambda(X_i'\beta_0))^2 + m(X_i)(1 - m(X_i)))\right],$$

where we have used $(Y_i - \Lambda_i)^2 = (Y_i - m_i + m_i - \Lambda_i)^2$, taken conditional expectations given X_i, and used $\mathbb{E}[(Y_i - m_i)(m_i - \Lambda_i) \mid X_i] = 0$.

The first term is the squared *specification error* and is generally nonzero. So $\Omega \neq Q$ – the information matrix equality fails. The asymptotic variance is then

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, Q^{-1}\Omega Q^{-1}),$$

and the simple Q^{-1} formula understates (or overstates) the true variance. Valid inference requires the **robust sandwich** estimator $\hat{V} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}$ where $\hat{\Omega}$ uses the empirical squared residuals $(Y_i - \Lambda(X_i'\hat{\beta}))^2$ rather than the model-implied $\Lambda(1 - \Lambda)$.

Problem 3: Consider a probit model with a single regressor: $\mathbb{P}(Y_i = 1 \mid X_i) = \Phi(\beta_0 + \beta_1 X_i)$, with $X_i \sim \mathcal{N}(0,1)$ and $(\beta_0, \beta_1) = (0,1)$. You wish to test $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$.

a. Write down the Wald, likelihood ratio, and score test statistics. State explicitly what each requires you to compute.

Solution

Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ denote the unrestricted MLE and $\tilde{\beta} = (\tilde{\beta}_0, 1)'$ the restricted MLE that imposes $\beta_1 = 1$.

Wald. Uses the unrestricted MLE only:

$$W = \frac{(\hat{\beta}_1 - 1)^2}{\widehat{\mathrm{Var}}(\hat{\beta}_1)},$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_1)$ is the $(2,2)$ element of $\hat{V}/n$, with $\hat{V} = \hat{Q}^{-1}$ (or the sandwich form under misspecification). Requires: estimate the unrestricted model and extract the variance of $\hat{\beta}_1$.

Likelihood ratio. Uses both restricted and unrestricted MLEs:

$$LR = 2(\ell_n(\hat{\beta}) - \ell_n(\tilde{\beta})).$$

Requires: estimate *both* models and take the difference of the maximized log-likelihoods.

Score (Lagrange multiplier). Uses the restricted MLE only:

$$LM = \frac{1}{n} S_n(\tilde{\beta})' \, \tilde{Q}^{-1} \, S_n(\tilde{\beta}),$$

where $\tilde{Q}$ is the negative expected Hessian evaluated at $\tilde{\beta}$. Requires: estimate the restricted model and compute the score and Hessian at the restricted estimate.

b. Under H_0, all three statistics are asymptotically equivalent and converge to a χ^2_1 distribution. Briefly explain *why* they are asymptotically equivalent, and give one practical reason you might prefer one over the others.

Solution

The three statistics are asymptotically equivalent because they all measure the same underlying object – the squared distance between the unrestricted and restricted parameter, scaled by the curvature of the log-likelihood at the optimum – but from different vantage points. A second-order Taylor expansion of ℓ_n around $\hat{\beta}$ gives

$$\ell_n(\tilde{\beta}) \approx \ell_n(\hat{\beta}) - \frac{1}{2}(\hat{\beta} - \tilde{\beta})' \mathcal{H}_n(\hat{\beta})(\hat{\beta} - \tilde{\beta}).$$

Substituting this into $LR = 2(\ell_n(\widehat{\beta}) - \ell_n(\widetilde{\beta}))$ yields exactly the Wald form $(\widehat{\beta} - \widetilde{\beta})' \mathcal{H}_n (\widehat{\beta} - \widetilde{\beta})$. A similar expansion of the score around $\widehat{\beta}$, using $S_n(\widehat{\beta}) = 0$, gives the equivalence with the LM statistic. All three converge to χ^2_1 under H_0 because $\sqrt{n}(\widehat{\beta}_1 - 1) \xrightarrow{d} \mathcal{N}(0, V_{11})$ and the square of a standard normal is χ^2_1.

Practical preferences: The Wald test is computationally cheapest if you have already estimated the unrestricted model – which is the typical workflow. The LR test is invariant to nonlinear reparameterizations of the parameter (Wald is not), so it is preferred for tests of nonlinear hypotheses. The LM test is useful when the restricted model is much easier to estimate than the unrestricted one, e.g., when testing whether to add a large block of regressors.

Problem 4: A researcher fits both a linear probability model (LPM) and a logit model to the same binary outcome data and reports that the LPM coefficient on a regressor X_j is 0.04, while the logit coefficient is 0.18. The researcher concludes that the logit model finds "a much larger effect."

a. Explain why this comparison is misleading. What is the correct way to compare effect magnitudes across the two models?

Solution

The comparison is misleading because the two coefficients live on different scales. The LPM coefficient is **already** a marginal effect on the probability scale: $\partial \mathbb{P}(Y = 1 \mid X)/\partial X_j = \beta_j^{\text{LPM}} = 0.04$. The logit coefficient is on the **log-odds** scale and must be transformed before it can be compared:

$$\frac{\partial \mathbb{P}(Y = 1 \mid X)}{\partial X_j} = \beta_j^{\text{logit}} \cdot \Lambda(X'\beta)(1 - \Lambda(X'\beta)).$$

Since $\Lambda(1 - \Lambda) \leq 1/4$, the logit marginal effect is at most $0.18 \times 0.25 = 0.045$, which is essentially identical to the LPM estimate. The "much larger effect" is an artifact of comparing apples to oranges.

The correct comparison is between the LPM coefficient and the logit **average marginal effect** (AME):

$$\widehat{\text{AME}}_j = \widehat{\beta}_j^{\text{logit}} \cdot \frac{1}{n} \sum_{i=1}^{n} \Lambda(X_i'\widehat{\beta})(1 - \Lambda(X_i'\widehat{\beta})).$$

LPM coefficients and logit/probit AMEs are typically very close in practice – this is the well-known empirical regularity behind the popularity of LPM as a quick approximation.

b. Suppose the LPM produces fitted probabilities $\widehat{P}_i$ in the range $[-0.05, 0.92]$. State two reasons why a binary choice researcher might still prefer logit despite the closeness of average marginal effects.

Solution

Reason 1: Boundary violations. The fitted values $[-0.05, 0.92]$ contain negative probabilities, which are nonsensical and signal that the linear functional form is a poor approximation in part of the covariate space. Logit guarantees $\widehat{P}_i \in (0, 1)$ by construction. Negative fitted values are particularly problematic when the goal is *prediction* (e.g., scoring a loan applicant) rather than estimating an average marginal effect.

Reason 2: Heteroskedasticity and efficiency. The LPM error has conditional variance $P(X)(1 - P(X))$, which depends on X. Although robust standard errors fix the inference problem, OLS is not the efficient estimator under heteroskedasticity – the logit MLE is. In samples where the heteroskedasticity is severe (probabilities near 0 or 1 for some observations), logit can be meaningfully more precise. This matters less for very large samples where both estimators are precise enough that the efficiency gain is academic.

Problem 5: A researcher estimates a logit model for the probability that a household owns its home, with regressors *income* (in tens of thousands of dollars) and *college* (a dummy equal to 1 if the household head has a college degree). The estimated coefficients are

$$\hat{\beta}_{\text{income}} = 0.25, \qquad \hat{\beta}_{\text{college}} = 0.80, \qquad \hat{\beta}_0 = -1.50.$$

a. Derive the relationship between the logit coefficients and the **odds ratio** $\mathbb{P}(Y = 1 \mid X)/\mathbb{P}(Y = 0 \mid X)$. Then interpret $\hat{\beta}_{\text{income}}$ and $\hat{\beta}_{\text{college}}$ on the odds-ratio scale.

Solution

Start from the logit specification $\mathbb{P}(Y = 1 \mid X) = \Lambda(X'\beta) = \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$. The complementary probability is $\mathbb{P}(Y = 0 \mid X) = \frac{1}{1+\exp(X'\beta)}$. Taking the ratio, the $1 + \exp(X'\beta)$ denominators cancel:

$$\frac{\mathbb{P}(Y = 1 \mid X)}{\mathbb{P}(Y = 0 \mid X)} = \exp(X'\beta).$$

This is the **odds** of $Y = 1$. Taking logs gives the **log-odds** (or **logit**):

$$\log\left(\frac{\mathbb{P}(Y = 1 \mid X)}{\mathbb{P}(Y = 0 \mid X)}\right) = X'\beta.$$

This is why the model is called "logit" – it is linear in the log-odds. Each coefficient β_j is the change in log-odds for a one-unit change in X_j, holding the other regressors fixed.

For multiplicative interpretation, exponentiate: increasing X_j by one unit multiplies the odds by $\exp(\beta_j)$. Applied to the estimates:

- $\exp(\hat{\beta}_{\text{income}}) = \exp(0.25) \approx 1.28$. A \$10,000 increase in income multiplies the odds of homeownership by roughly **1.28**, a 28% increase in the odds (not in the probability).

- $\exp(\hat{\beta}_{\text{college}}) = \exp(0.80) \approx 2.23$. A college-educated household head has odds of homeownership roughly **2.23 times** those of an otherwise-identical non-college household – more than double the odds.

b. A student looks at part (a) and concludes: "So a college degree more than doubles the *probability* of homeownership." Explain carefully why this conclusion is wrong, and compute the correct marginal effect of *college* on the probability of homeownership for a household with *income* = 5 (i.e., \$50,000).

Solution

The student is confusing **odds** with **probability**. Doubling the odds is not the same as doubling the probability – the two only coincide when probabilities are very small (because $p/(1-p) \approx p$ for p near zero). When probabilities are moderate or high, a doubling of odds corresponds to a much smaller proportional change in probability. A simple example: going from $p = 0.5$ (odds = 1) to $p = 0.667$ (odds = 2) doubles the odds but only raises the probability by 17 percentage points.

To compute the correct probability effect of *college* at *income* = 5, evaluate the predicted probability at both values of *college*. The linear index for a non-college household is

$$X'\hat{\beta} = -1.50 + 0.25 \times 5 + 0.80 \times 0 = -0.25,$$

giving

$$\Lambda(-0.25) = \frac{1}{1+\exp(0.25)} \approx \frac{1}{1+1.284} \approx 0.438.$$

For a college household at the same income:

$$X'\hat{\beta} = -1.50 + 0.25 \times 5 + 0.80 \times 1 = 0.55,$$

giving

$$\Lambda(0.55) = \frac{1}{1+\exp(-0.55)} \approx \frac{1}{1+0.577} \approx 0.634.$$

The discrete probability effect of a college degree at *income* = 5 is therefore $0.634 - 0.438 \approx 0.196$, or about a **20 percentage point** increase in the probability of homeownership – not a doubling. The probability rises from 44% to 63%.

The general lesson: logit coefficients have a clean multiplicative interpretation on the odds-ratio scale, but if you want to communicate effects in probability units (which is usually what an audience cares about), you must compute predicted probabilities or marginal effects directly, and these will depend on where in the covariate space you evaluate them.